A Novel Data Classification with Anonymity Method for Privacy Preserving in Medical Data Mining

Dr.C.Senthilkumar
Associate Professor, Department of Computer Science,
Erode Arts & Science College (Autonomous), Erode, Tamil Nadu, India.
Email: csincseasc@gmail.com
P.Kalaiyarasi
M.Phil (Research scholar), Department of Computer Science,
Erode Arts & Science College (Autonomous), Erode, Tamil Nadu, India.
Email: kalaipavi6@gmail.com

Abstract— Data mining is the process of extracting interesting patterns or knowledge from huge amount of data. The privacy preserving in data mining comes into picture for security. K-Anonymity is one of the easy and efficient techniques to achieve privacy preserving for sensitive data in many data publishing applications. In K-anonymity techniques, all tuples of releasing database are generalized to make it anonymized which leads to reduce the data utility and more information loss of publishing table. To overcome those problems, it needs to propose a model is called Novel Sensitive Class Based Anonymity Method (NSCBA).The proposed the method classifies sensitive attributes like high sensitive and low sensitive depending upon the sensitive values. Experiment results on the SPARCS medical data sets show the proposed methods not only can improve the accuracy of the publishing data, but also can preserve privacy, then can increase the data utility and minimum information loss and also provide privacy with the implementation of ASP.NET.

Keywords—K-Anonymity, Privacy Preserving, NSCBA, SPARCS, Sensitive data.

1. INTRODUCTION

The tremendous growth in Information and Communications technology increases the need for electronic data to be stored and shared securely. The huge amount of data, if publicly available can be utilized for many research purposes. Data Mining can be one of the technologies used to extract knowledge from massive collection of data. On the other hand, being published, the sensitive information about individuals may be disclosed which creates ethical or privacy issues [1].

Due to privacy issues many individuals are reluctant to share their data to the public which leads to data unavailability. Thus, privacy should be an important concern in the field of Data Mining. Privacy Preserving Data Mining (PPDM) is becoming a popular research area to address various privacy issues. Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources and put together so that they can be analyzed [2]. This is not only data mining, but also is used for the result of the preparation of data before the purposes of the data analysis.

The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous. Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community. The increasing ability to track and collect large amounts of data with the use of current hardware technology that has led to an interest in the development of data mining algorithms which preserve user privacy. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records.

Various data mining techniques, it allows sharing of large amount of privacy sensitive data for analysis purposes. One of the major problems of privacy preserving data mining is the abundant availability of personal data. This paper is going to discuss about introduction for privacy preserving in data mining in Section I. The related works on anonymization approaches regarding hidden data in Section II. In Section III Problem definition is decided. The proposal model is discussed in Section IV. The Experimental result are presented in Section V. Finally the conclusion for the research work objectives is concluded in Section VI.

2. RELATED WORKS

In recent years, many algorithms have been proposed for implementing k-anonymity via generalization and suppression. Samarati [4] presented an algorithm that exploits a binary search on the domain generalization hierarchy to find minimal k-anonymous table. Model such as l-diversity proposed in 2006 by A. Machanavajjhala [5] solve k-anonymity problem. It tries to put constraints on minimum number of distinct values seen within a equivalence class for any sensitive attribute.

S. Venkatasubramanian in 2007 [3] developed a model called t-closeness which was introduced to overcome attacks possible on l-diversity like similarity attack. R. Wong, J. Li, A. Fu, K. Wang [7] proposed an (α, k)-anonymity model to protect both identifications and relationships to sensitive

information in data that were proposed in the literature in order to deal with the problem of k- anonymity.

Bayardo and Agrawal [9] published an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal k- anonymous table. Fung et al. [8] designed a top-down approach to make a table satisfied k-anonymous .LeFevre [6] described an algorithm that uses a bottom-up technique. However the traditional k-anonymity models consider that the all values of the sensitive attributes are sensitive and need to be protected.

Vaidya[10] developed the methods for privacy-preserving association rule mining. The perturbation approach restricts data services to learn or recover precise records. This restrictions leads to some challenges. As this method does not reconstruct the original data values excepting distribution, so new algorithms require to be developed for reconstructed distributions to perform mining of the underlying data.

Jisha Jose Panackal[11] discussedthe approaches in different way and is applicable when data can be disclosed beyond the control of the data collection process. If the data is distributed across multiple sites which are legally prohibited from sharing their collections with each other, it is still possible to construct a data mining model. The paper [12] is illustrated this scenario. They proposed a cryptographic protocol based on decision-tree classification on horizontally partitioned databases.

Vaidya and Clifton first analyzed how secure association rule mining can be done for vertically partitioned data by extending the Apriori algorithm. Du and Zhan [16] developed a solution for constructing ID3 on vertically partitioned data between two parties. Clifton [14] presented a Naive Bayes classifier for privacy preservation on vertically partitioned data and [15] proposed the first method for clustering over vertically partitioned data. All these methods are almost based on the special encryption protocol known as Secure Multiparty Computation (SMC) technology. The SMC originated with Yao's Millionaires' problem [13].

In fact, the values which will breach individual's privacy are in the minority of the whole sensitive attribute dataset. The previous models lead to excessively generalize and more information loss in publishing data. The work presented in this paper mainly considers the tuples which are really sensitive and need to be preserving the privacy of individual are only generalized and anonymized.

3. PROBLEM DEFINITION

The goal of data mining is to extract hidden or useful unknown interesting rules or patterns from databases. The main objective of privacy preserving data mining is to hide certain confidential data so that they cannot be discovered through data mining techniques. In research work survey analysis, the problem specification is defined according to the collected data. Based on collected data, our problem may be extended to another area or applications. After defining problem, the components (or) sub modules of the problem are analyzed. The main problem is to secure the private data and avoid the information loss and use maximum data storage. The needed database is designed for updating of the available data.

Privacy – To provide the individual data privacy by generalization in such a way that re-identification cannot be possible.

Data utility - The goal is to eliminate the privacy breach (how much an adversary learn from the published data) and increase utility (accuracy of data mining task) of a released database. This is achieved by generalizing quasi-identifiers of only those tuples having high sensitive attribute values.

Minimum information loss – The loss of information is minimized by giving sensitivity level for sensitive attribute values, and tuples which belongs to high sensitive level are only generalized rest of the tuples are released as it.

Basic Notation

Let T{K1,K2.. ,Kj Q1,Q2,..,Qp,S } be a table. For example, T is a medical dataset. Let Q1,…,QP denote the quasi-identifier specified by the application (administrator). Let S denotes the sensitive attribute.

Definition 1: (Quasi-identifier)

A set of non-sensitive attributes {Q1,…,Qp} of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify (can be called as candidate key) at least one individual in the general population.

Definition 2: (K-Anonymity)

A table T satisfies k- anonymity if for every tuple t of T there exists (k-1) other tuples ti1, ti2, tik-1 $\in$ T such that t[F] = ti1 [F] =ti2 [F] = … = tik−1 [F] for all F$\in$ QI.

Definition 3: (Sensitive-values Set)

A Set A consists of values which the user selects as most sensitive values from set S which denote by A.

Definition 4 : (Sensitive tuple)

Let t $\in$ T, if t[S] $\in$A, t is called as sensitive tuple.

4. PROPOSED METHOD FOR NOVEL SENSITIVE CLASS BASED ANONYMITY METHODS(NSCBA)

K-anonymity model is introduced to protect sensitive attributes from interlopers where sensitive attribute is an attribute whose value for some particular individual must be kept secret from people who have no direct access to the original data. Data publisher needs to prevent privacy disclosure which means someone can simply attack link the publish table T and at least know the individuals suffer from some kinds of privacy disease. This phenomenon is a kind of privacy disclosure in data mining set. Information disclosure is of three types:

Identity Disclosure: An individual is linked to a particular record in the published data.

Attribute Disclosure: Sensitive attribute information of an individual disclosed.

Membership Disclosure: Information about whether an individual's record is in the published data or not disclosed.

They are to be generalized and the publishing data lost a lot of useful information. The kernel idea is to protect individual's privacy as well as only the high sensitive tuples should be generalized with a satisfied parameter K. The other tuples should not be generalized and can be published directly.

K-Anonymity is known as representative anonymization technique. To identify records uniquely, it

considers quasi identifiers which can be used in conjunction with public records.

Data publishers have to face problem when multiple sensitive attributes are present in records. The traditional K-Anonymity method takes all tuples as sensitive. The new proposed method is called Novel Sensitive Class Based Anonymity Method (NSCBA).

The method can classify the disease into two type high, low it depends upon the sensitive values.

High Sensitive values: A set of sensitive attribute values H= {s1, s2…sn} that are highly sensitive like HIV, Cancer.

Low Sensitive Values: A set of sensitive attribute values L= {s1, s2…sk) that are low sensitive like Flu, Viral infection.

// Novel SCBA Algorithm**//**

Input -Table T , set of Quazi identifier Q,

Output-Anonymized table T*

Step 1: Select Input table and Q is set of quazi-identifier attributes

Step 2: Select sensitive attribute S.

Step 3: Classify sensitive values in two classes H and L.

Step 4: Identify the Quazi attribute in high sensitive value.

Step5: For each tuple whose sensitive value belongs to set H i.e. if t[S] ∈H then move all these tuples to Table T1, and apply generalization on Quazi attribute so that tuples get anonymized.

Step 6: If t[S] ∈L then move all these tuples to Table T2.

Step 7: Append rows of table T1, T2.
T*=T1+T2 which is table ready to release.

Step 8: End process.

The applying SCBA algorithm, Sensitive values HIV and Cancer are selected as High sensitive value and tuples belonging to those values are moved to Table and generalization is applied on quazi attributes Zip code, Age and Sex to anonymized those tuples. Sensitive values like Flu and Headache etc., are selected as Low sensitive values and they are stored and released as it is.

The proposed system operates in five phases

1) Identify Key attribute: In the given dataset the proposed algorithm take a one tuple as key attribute using the key attribute that can be easily to find someone. First up all, algorithm finds key attribute for (example Name, country).

2) Identify Quazi-attribute: Second stage, algorithm can find out the quazi attribute like Name, Zip code, Sex. The quazi attribute is essential for making anonymization.

3) Identify Sensitive attribute: Sensitive attribute can be chosen randomly. In medical dataset selected diseases are the sensitive attribute.

4) Classification: Classification is applied the sensitive attribute that can classify the disease into two types high and low, it depends upon the sensitive values.

5) Anonymization: Anonymization includes two types of techniques such as suppression and generalization. Suppression is performed on quazi attributes are replaced by * (example name, address). Generalization is performed on quazi attributes that are replaced by border gateway (example if age is 25 then replace by >20).

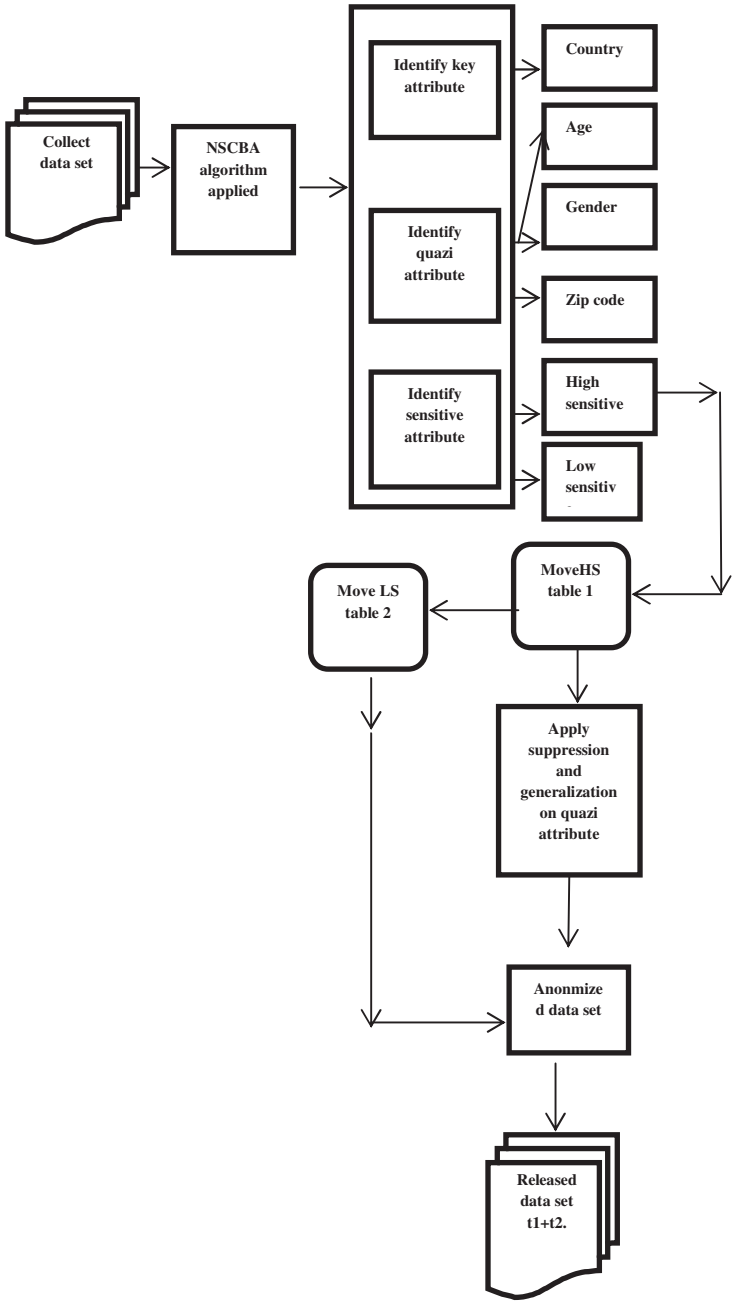

Fig 4.1 process flow diagram for proposed NSCBA algorithm

Algorithm For Hidden the Medical Data

Using the suppression and generalization technique are used for hidden the original data into replace some values. In suppression method some attribute are replaced by *, then in generalization method the attributes are replaced by border gateway. The algorithm performs hiding the high sensitive attributes in database. The attributes of tables are classified into three classes. Algorithm identifies the unique attribute such as id, key value. Algorithm identifies the common attributes that are publicly available in all records as quazi attributes. Sensitive attributes are the attributes which are need to be protected.

// **Algorithm for Hidden Medical Data//**

Input- A data set D, quazi-identifier attributes Q, high sensitive attribute HS,

Output- Releasing table d*.

Step1: Create classification for the cluster k= Africa, Australia, Russian, Indian, China, America.

Step2: For each one clusters do step3-step 7 until last one.

Step3: Classify the high sensitive data and low sensitive data.

Step4: In high sensitive data select quazi attributes and sensitive attributes.

Step5: Hidden quazi attributes in records are using

(Gender, Zip code, Age) field are stored.

Step6: Display the low sensitive data without anonymized.

Step7: Repeat step5 until the last cluster.

Step8: Stop the process.

The selection of sensitive attributes is important because, there is a need to anonymized only the most sensitive data to avoid the overhead and to increase the data utility. After identifying all of the attributes, suppression is applied only to the key attribute. This algorithm suppression it makes with the help of special symbols, and generalization makes the border gateway of the attribute, to make the hidden in selective attribute to get anonymized.

5. EXPRIMENTAL RESULTS

Medical database named SPARCS including about 10, 48,576 medical data form various countries is consider as dataset. The SPARCS is a comprehensive data reporting system established in 1979 as a result of cooperation between the Health Care Industry and Government. The medical data set, sub data set contains the 10, 48,576 record and 38 attributes. Preprocessing is applied the original dataset proposed method to get 30,000 records in 9 attributes for the proposed method.

Preprocessing from data set can take six country namely Africa, India, Russian America, Australia, China. After proposed method classifies the country wise, each country is having 5000 records. In each and every country has a high and low sensitive disease. The research model classifies the disease depending up on the sensitive value .After that collected data goes to hidden the high sensitive quazi attributes.

Table 5.1 Data set Information

Data set	SPARCS(medical data)
No of records	30,000
No of attributes	9

In data set each and every country has one key attribute, quazi attribute and sensitive attributes. The sensitive attributes are considered to classify the sensitive value.

The following Fig 5.1&5.1.1 show 30,000 extracted data set information from the original data set.

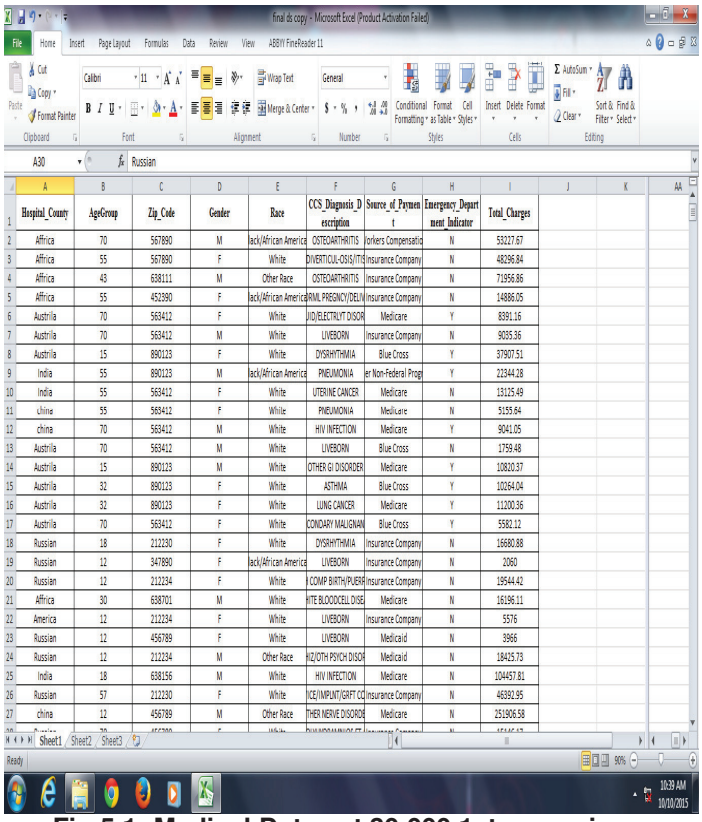

Fig 5.1: Medical Data set 30,000 1st page view

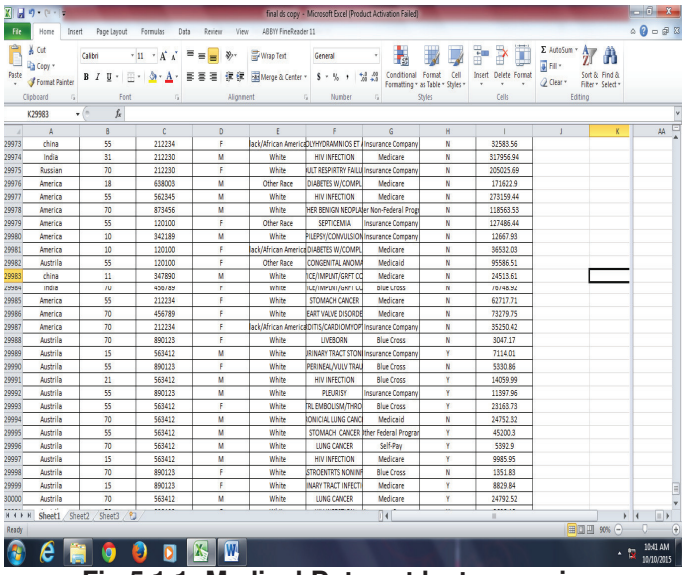

Fig 5.1.1: Medical Data set last page view

The above screen shows the data set for medical data available in excel format. The data set indicates the information about the various countries. Each country have unique table. Experiments are conducted using ASP.NET. Many different methods for measuring the performance of a system have been created and used by researcher.

The proposed method is used for classification and clustering. The Encryption and Decryption are used in the anonymity methods like suppression and generalization. The first method classifies the each and every country. Finally, cluster for the similar objects are obtained and shown in fig 5.2&5.3.

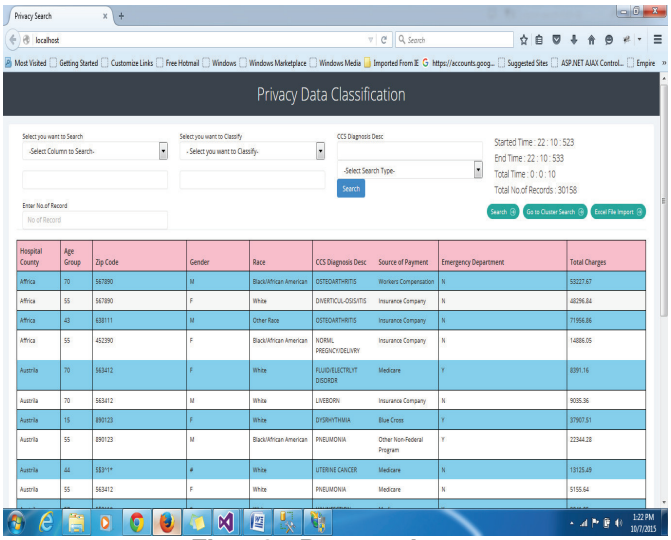

Fig 5.2: Data set import

Fig 5.3: Classification data for Africa country

In figure 5.4, the obtained low sensitve data for Africa are indicated.

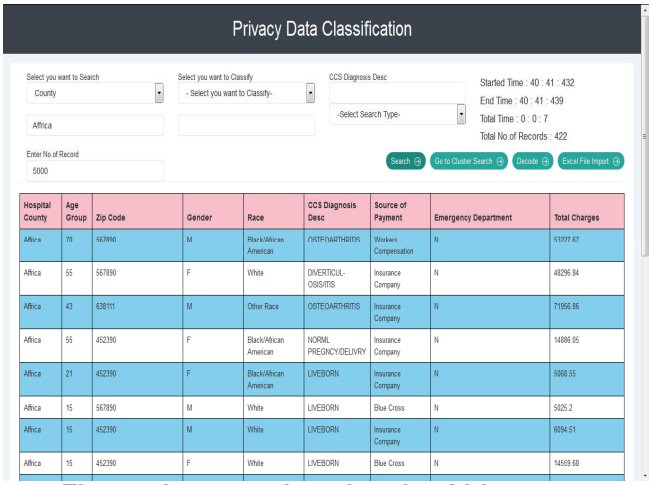

Fig 5.4: Low sensitve data for Africa country

Similarly,high sensitve data for Africa country are presented in figure 5.5.

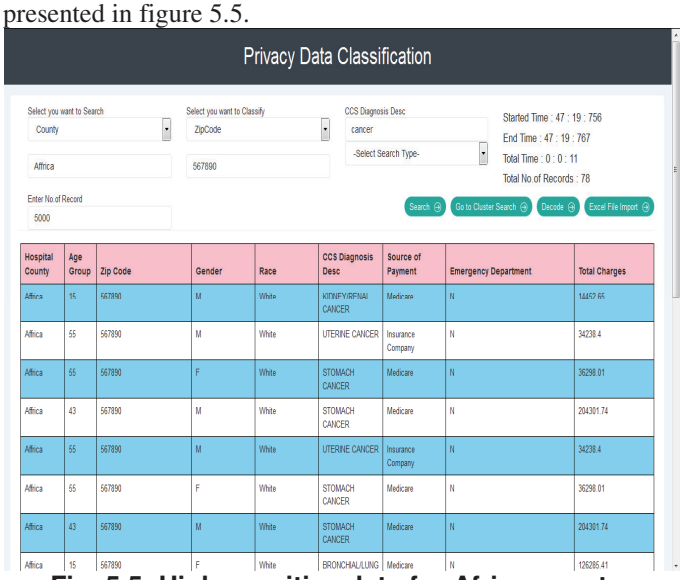

Fig 5.5: High sensitive data for Africa country

Fig 5.6,Encode table is generated with the help of encryption fom for the above data set.

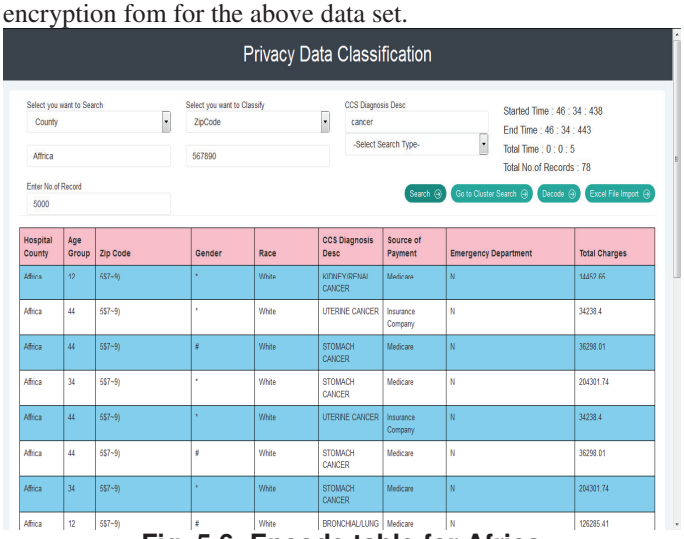

Fig 5.6: Encode table for Africa

Similarly, the coressponding Decode is performed on the data set and is presented in the fig 5.7.

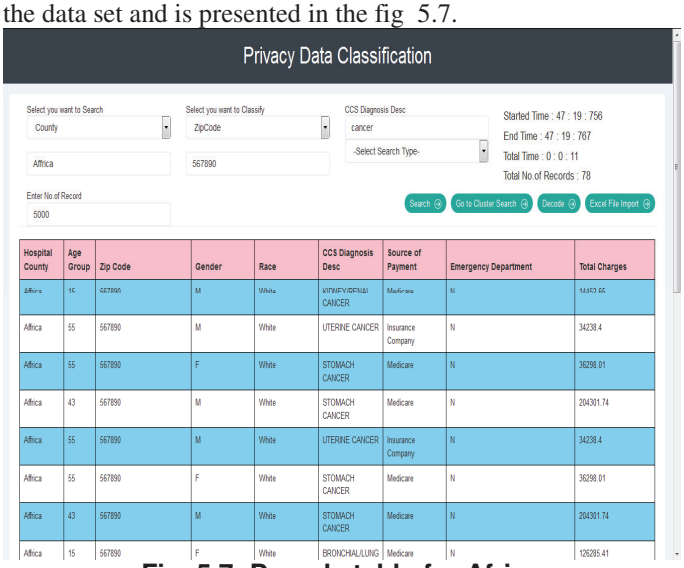

Fig 5.7: Decode table for Africa

The proposed evolution method consists of three stages. First classification is performed the country wise. The proposed algorithm classifies the country that has high sensitive and low sensitive data. The clustering algorithm is performed the group of similar objects. Similar objects contain the same cluster. High sensitive attributes are hidden the quazi attributes using suppression and generalization methods. The encryptions of those attribute can extract (or) decrypt the original attribute values. In each and every sensitive value depends upon the sensitive attributes. In India country, the proposed method can take zip code to find out which area is most affected in high sensitive diseases.

Table 5.2 Classification for six countries

Database size	C1		C2		C3		C4		C5		C6	
5,000	N.R	T	N.R	T	N.R	T	N.R	T	N.R	T	N.R	T
	657	9	877	9	754	8	1303	10	672	10	737	6
10,000	1700	9	1317	14	1528	11	2130	8	1600	9	1725	8
15,000	2898	10	2380	7	2700	8	2275	17	2702	12	2045	9
20,000	3853	10	3711	9	4073	8	2426	9	3302	8	2635	7
25,000	4623	10	4159	8	4870	9	3293	8	4162	9	3893	8
30,000	5209	7	4484	14	5520	8	4365	7	4939	7	4483	9

The above Table 5.2 shows the classification for the six countries. In each and every country contains the some attribute values. This table focuses only time for classification of each country.

Table 5.3 Comparison Search time in Medical data set

Record size	Methods		
	Sequential search(time ms)	Binary search (time ms)	Cluster search (time ms)
5,000	10	7	6
10,000	9	6	4
15,000	6	3	2
20,000	9	8	6
25,000	7	6	5
30,000	4	3	2

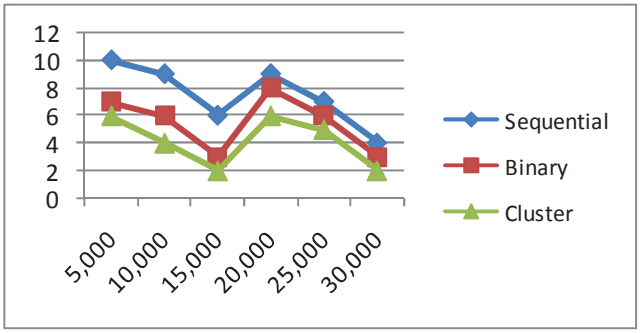

Fig 5.8: comparison search time

The comparison for three searching methods is tested in successfully. Sequential search takes long time while compare to other two search methods. Binary search is reducing the time compare than the sequential search. Finally cluster search is taken a minimum time to search the data in data base. Cluster search gives better result which compared to sequential and binary search. The result is presented in the table 5.3 from fig 5.8; it is observed that the proposed cluster search ensures the research objectives.

It is obvious observed that the proposed method compares the three Anonymity algorithm which one is best retrieval and hidden the medical records. Process time is calculated for hidden the quazi attributes in the medical data.

Table 5.2 Time calculation for hidden the attributes

Data base size	Existing Time to hide k-anonymity	Existing Time to hide (ARH)MS	Proposed Time to hide NSCBA	Total % to hide improved
10000	21	19	18	31
15000	44	29	28	34
20000	71	42	41	41
25000	85	49	48	43
30000	112	61	60	47

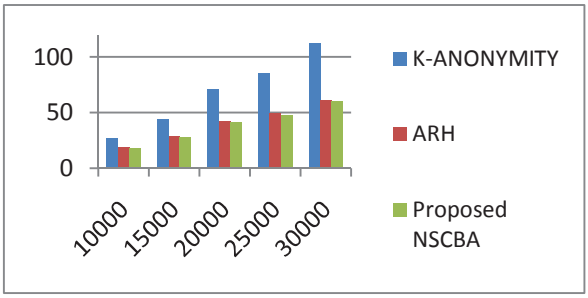

Fig 5.9Time calculation for hidden attributes

It is noted that there are three anonymity algorithm namely K-Anonymity, Association Rule Hiding and Novel Sensitivity Class Base Anonymity method. The above all three algorithms are anonymity methods. The both K-Anonymity and Association Rule Hiding methods can hide all sensitive records at the database. Those methods can provide information loss and unsecured records.

In NSCBA methods to hide the high sensitive data like Hiv, Cancer relevant quazi attributes are hidden using suppression and generalization. These methods can provide minimum information loss and utility of data is high and also can reduce the storage size. Comparison of those three algorithms for NSCBA gives better hidden time for the above algorithms.

6. CONCLUSION

K-Anonymity protects against the identity disclosure, it does not provide protection against homogeneity attack and background knowledge attack. As concluding remark, there are main issues or threats in traditional K-Anonymity privacy preserving algorithms. In existing algorithm consider all of sensitive attribute values at same level and apply generalization on all, this leads to some issues like, Information Loss, Data Utility and Privacy. The newly presented research proposed method Anonymity method in this paper rectified these issues based on classification sensitive attribute by which information is reduced. This paper presented a new K-Anonymity model based on sensitive attributes, by which information loss is reduced. Only sensitive attributes are anonymized by this method, so data utility is increased. Excessive generalization and suppression leads to the reduction of data utility and more information loss of publishing data. This is a secure algorithm to maintain usability and privacy of data sets. The proposed model can be extended for the following research domain

- Cloud Storage
- Large volume data such as telephone directory, internet Email user.
- With the help of fuzzy logic, this work may be modified for automatic systems.

7. REFERENCES

[1] Jisha Jose Panackal and AnithaS.Pillai, "Privacy Preserving Data Mining an Extensive Survey", Association of Computer Electronics and Electrical Engineers,pp.297-304,2013.

[2] Seema Kedar,Sneha Dawdle and WankhadeVaibhav," Privacy Preserving Data Mining", International Journal of Advanced Researching Computer and Communication Engineering Vol.2, Issue 4, pp.1677-1680 , April 2013.

[3] N. Li, T. Li, S. Venkatasubramanian. t-.Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007:106-115.

[4] Samarati. Protecting respondents‟ identities .in micro data...release. IEEE Transactions on .Knowledge and Data Engineering, (6):10101027.2001.

[5] A.Machanavajjhala,J.Gehrke,Kifer,.M.Venkitasubramaniam,“l-Diversity: Privacy beyond k-anonymity” In: Proceedings of the IEEE ICDE 2006.

[6] K.LeFevre, D.DeWitt, R.Ramakrishnan. Incognito: Efficient full domain k-anonymity Proceedings of the ACM SIGMOD International Conference on Management of Data Baltimore Maryland ,2005:49-60.

[7] R.Wong, J.Li, A. Fu, K. Wang. (α, k)-anonymity: an enhanced ..k-anonymity model For privacy preserving data publishing .KDD 2006:754-759.

[8] B.Fung, K. Wang, P. Yu. Top-down ..specialization for information Conference on Data Engineering (ICDE05), pp:205-216.

[9] R.Bayardo and R. Agrawal. Data privacy through optimal k-anonymity. In. Proceedings of the 21st International conference on Data Engineering (ICDE),.pp:217-228,Tokyo,Japan,2005.

[10] I.Ioannidis, A.Grama, M.J.Atallah, “A Secure Protocol for Computing Dot-Products in Clustered and Distributed..Environments”, In Proceedings of the 31st International Conference on Parallel Processing, pp.379-384, 2002.

[11] Geetha Jagannathan, Rebecca N.Wright, “Privacy-Preserving..Imputation of Missing Data”, Data & Knowledge Engineering, .Elsevier 2008.

[12] Yao,C.Andrew,“ How to Generate and Exchange Secrets”, In proceedings of the 27th IEEE Symposium on Foundation of Computer Science, pp.162-167, 1986.

[13] J.Vaidya, C. Clifton, “Privacy Preserving Naive Bayes Classifier for Vertically Partitioned Data”, In Proceedings of the 2004 SIAM International Conference on Data Mining, pp.522–526, 2004.

[14] J. Vaidya, C. Clifton, “Privacy-Preserving k-Means Clustering over Vertically Partitioned Data”, In Proceedings ofthe 9th. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.206–215, 2003.

[15] W.L. Du, Z.J. Zhan, “Building Decision Tree Classifier on .Private Data”, In Proceedings of the IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, pp.1-8, 2002.

[16] Rao, R.B, Krishnan, S. and Niculescu,R.S, "Data Mining for Improved Cardiac Care", SIGKDD Explorations Volume8.

[17] Sachin Janbandhu and S.M.Chaware, "Survey on Data Mining with Privacy Preservation", (IJCSIT) International Journal of Computer Science and Information Technologies, ISSN:0975, 4676, Vol. 5 (4),pp. 5279-5283, 2014

[18] N.S.Nithya, K.Duraiswamy and P.Gomathy, “A Survey on Clustering Techniques in Medical Diagnosis", International Journal of Computer Science Trends and Technology (IJCST), Volume1 Issue2, pp.17-22, Nov-Dec, 2013.

[19] E.Poovammal and M.Ponnavaikko, “An Improved Method for Privacy Preserving Data Mining”, IEEE International .Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.

[20] R. Adam and J.C. Wortman, "Security-Control Methods for Statistical Databases: A Comparative Study", ACM Computing Surveys, vol.21, no.4, pp. 515-556, 1989.

[21] Arun k. pujari, "Data mining Techniques" university Press, First Edition 2001.

[22] D.Kinoshenko, V.Mashtalir and E.Yegorova, "Clustering method for fast content- Based image retrieval" Computer Vision and Graphics,32,2006.

[23] Kusiak,A., Kernstine,K.H., Kern, J.A, McLaughlin, K.A., and Tseng,T.L., "Data mining: Case Studies". Proceeding of the Industrial Engineering Research 2000 Conference,Cleveland,Ohio,pp.1-7,21-23, 2000.